A Database-Driven and Machine Learning-Based Approach for Estimating Consolidation Parameters and Drainage Conditions Using CPTu Data

Iman Entezari, James Sharp, Dallas McGowan
ConeTec Group, Burnaby, Canada, ientezari@conetec.com

Jason DeJong
University of California Davis, Davis, United States

ABSTRACT: Piezocone penetration testing (CPTu) with pore pressure dissipation (PPD) measurements is a valuable tool for assessing drainage conditions and estimating the horizontal coefficient of consolidation (c_h) in soils. A key parameter in this process is t_{50}, the time required to reach 50% excess pore pressure dissipation, which is commonly measured to calculate c_h. However, dissipation tests are time-consuming and often limited to a few locations per site. To explore more efficient interpretation methods, this study applied a machine learning (ML) framework to predict t_{50} from early-time dissipation data and CPTu parameters. The study also examined the potential of ML to classify soils as drained, partially drained, or undrained. A database of 2,695 high-resolution dissipation tests from recent ConeTec projects was compiled to support the analysis. In addition, the data were used to evaluate and refine empirical frameworks, such as the time factor (T*) method commonly used for c_h estimation. While the results are preliminary, they demonstrate the potential of ML and database-driven approaches to support more efficient and informative CPTu dissipation test interpretation.

KEYWORDS: Piezocone penetration testing (CPTu), Pore pressure dissipation (PPD), Machine Learning (ML).

1 INTRODUCTION

Piezocone penetration testing (CPTu) is widely used for geotechnical characterization in tailings and soft ground engineering. Most interpretation frameworks are developed under the assumption that the soil response during penetration at the standard rate of 2.0 cm/s is either fully drained or fully undrained. However, many transitional soils, such as silts, silty clays, and clayey sands, exhibit partially drained behavior due to their intermediate permeability (Robertson, 2016; DeJong & Randolph, 2012). This deviation from idealized assumptions challenges the reliability of conventional interpretation methods for estimating strength and consolidation parameters in such materials.

The time to 50% excess pore pressure dissipation (t_{50}), derived from pore pressure dissipation (PPD) tests, is a widely used parameter for assessing drainage conditions and calculating the horizontal coefficient of consolidation (c_h), used in evaluating drainage and settlement behavior in soft, fine-grained soils. However, dissipation testing is time-intensive and typically limited to a small subset of CPTu soundings due to operational constraints. This limits the spatial resolution of drainage assessments and c_h estimates across a site.

This paper explores the use of machine learning (ML) and database-driven approaches to improve the estimation of t_{50} and enable more effective classification of soil drainage behavior. Specifically, it investigates the potential to predict t_{50} in real time using standard CPTu parameters such as tip resistance (q_t), sleeve friction (f_s), pore pressure (u_2), and depth, thereby eliminating the need to pause for conventional dissipation testing. Real-time estimation of t_{50}, and by extension c_h, has the potential to support adjustment of the penetration rate to achieve drained or undrained CPTu soundings. The study also examines the feasibility of estimating t_{50} from short-duration PPD tests, which could improve operational efficiency without substantively sacrificing accuracy. Furthermore, ML models are developed to classify the drainage condition (i.e., drained, undrained, or partially drained) based on both basic CPTu data and short PPD records. In addition, a database-driven approach is employed to empirically estimate the time factor at various dissipation levels, with comparisons made to the theoretical values proposed by Teh & Houlsby (1991).

2 DATASET DESCRIPTION

To support the objectives of this study, a comprehensive database of PPD tests conducted by ConeTec between 2020 and 2023 was compiled. PPD tests acquired within this period included high time-resolution dissipation data (on the order of sub-seconds) enabled by ConeTec's digital cone and data acquisition systems. Such fine temporal resolution is particularly important for accurately capturing t_{50} in dissipation tests where pore pressure dissipates rapidly. The dataset was filtered to retain tests conducted with 15 cm² cones and reaching at least 50% dissipation. It primarily consisted of dissipation tests with monotonically decaying pore pressure responses. However, for the fraction of tests exhibiting a brief initial dilatory response, t_{50} was calculated relative to the recorded maximum pore pressure (u_{max}) and referenced to the time at which u_{max} occurred. Dissipation times corresponding to other degrees (e.g., 20–80%) were also calculated. For each test, the associated CPTu parameters (q_t, f_s, u_2, and depth) were averaged over a 20 cm window centered at the test depth to create a paired CPTu–t_{50} dataset. The final dataset comprised 2,695 data points from 1,140 soundings. It included test results collected from a wide range of soil types, which were predominantly from North America and supplemented by additional sites in South America, Australia, Europe, and Asia. Table 1 lists the summary statistics of the dataset.

Table 1. Summary statistics of the CPTu–t_{50} dataset.

	Min	Max	Mean
t_{50} (s)	0.6	22026	392.93
q_t (kPa)	130.37	44984.13	3581.55
f_s (kPa)	0.75	1121.1	68.12
u_2 (kPa)	-61.53	5151.8	447.36
depth (m)	1.0	118.2	18.10

3 METHODOLOGY

3.1 *Estimation of t_{50} using ML*

To investigate the feasibility of estimating t_{50} through data-driven methods, two ML-based approaches were evaluated: (1) using basic CPTu parameters, and (2) incorporating results from short pore pressure dissipation (PPD) tests.

To develop a model for estimating t_{50} from CPTu parameters (q_t, f_s, u_2, and depth), an XGBoost (Chen & Guestrin, 2016) regression model was trained using the paired CPTu–t_{50} dataset. Since the dataset was highly skewed toward low t_{50} values, a log transformation was applied to the target variable to make its distribution more uniform and improve model performance. While XGBoost is relatively robust to target skewness due to its tree-based structure, the transformation can further stabilize training and enhance predictive accuracy. The dataset was randomly split into 72%, 13%, and 15% subsets for training (n = 1946), validation (n = 344), and testing (n = 405), respectively. The validation set was used for Bayesian optimization of hyperparameters, with root mean squared error (RMSE) as the scoring metric. After identifying the optimal parameters, the final model was retrained on the combined training and validation sets and then evaluated on the test set.

Separately, a Long Short-Term Memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997) was trained to predict t_{50} based on early-time pore pressure dissipation data. LSTM networks are well-suited to modeling temporal dependencies in sequential data and can effectively capture the characteristic decay patterns in dissipation curves. A partial dissipation level approach was tested, which utilizes the portion of the dissipation curve corresponding to 20% and 30% dissipation. This approach requires prior knowledge or estimation of the equilibrium pore pressure (u_{eq}) to accurately determine dissipation levels. To reduce variability between tests, each dissipation curve was normalized by first subtracting the u_{eq} from the measured pore pressure values, and then dividing by the difference between the maximum pore pressure and u_{eq}. This produced a dimensionless dissipation response ranging from 1 to 0. Additionally, the data was preprocessed so that each dissipation curve was resampled to have the same time resolution, enabling consistent input for the LSTM. As with the XGBoost model, a log transformation was applied to the t_{50} target, and the same training, validation, and test splits were used. The validation set was used to manually tune LSTM hyperparameters such as the number of layers and the number of hidden units per layer.

The final performance of each trained model was evaluated using the mean absolute percentage error (MAPE) on the test dataset. MAPE was selected to provide a percentage-based measure of prediction accuracy which measures the average relative difference between the predicted and true t_{50} values, making it a more appropriate choice for capturing proportional errors across a wide range of values.

3.2 *Classification of soil drainage condition using ML*

To assess the potential of ML modeling to classify soil drainage conditions, records in the compiled dataset were categorized as drained, partially drained, and undrained based on t_{50} thresholds of less than 5 seconds, between 5–75 seconds, and greater than 75 seconds, respectively (based on DeJong et al., 2012). While fully drained conditions are typically associated with near-instantaneous pore pressure dissipation (i.e., t_{50} approaching zero), the dataset contained very few tests with such rapid dissipation. As a result, the threshold for the drained category was selected to be 5 seconds to include a broader range of relatively fast-dissipation cases and improve sample representation across classes. The chosen thresholds therefore reflect a practical compromise between ideal definitions and the available data distribution.

To ensure a fair and balanced evaluation, 50 samples from each class were randomly selected for the validation set, and another 50 per class for the test set. The remaining records were reserved for training. As the overall dataset was imbalanced, with a disproportionately lower number of drained samples compared to the other two classes, Synthetic Minority Oversampling Technique (SMOTE) was used to address this imbalance in the training data (Chawla et al., 2002). SMOTE works by generating synthetic samples of the minority classes through interpolation between existing examples, improving the ability of the model to learn representative decision boundaries.

Two machine learning approaches were evaluated: 1) an XGBoost classifier trained on basic CPTu parameters, and 2) an LSTM neural network trained on 10 seconds of time-resolved pore pressure dissipation data. In both cases, the validation set was used to tune the hyperparameters of the models, and final performance was assessed on the test set using the accuracy metric, calculated as the proportion of correctly classified samples out of the total number of test samples.

3.3 *Determining T^* from a database approach*

To investigate a database approach for the empirical estimation of T^* values across different dissipation levels, the relationships between t_{50} and t_{20}, t_{30}, ..., t_{80} was analyzed using the compiled dataset. In this approach, t_{50} and T_{50} of 0.245 (per Teh & Houlsby, 1991) are used as reference points, and all other dissipation times and time factors are compared relative to them.

In practice, c_h is most commonly calculated using the time to 50% dissipation (t_{50}) and a corresponding T^* value of 0.245 for the u_2 filter position using the Equation (1) proposed by Teh & Houlsby (1991):

$$c_h = \frac{T^* r^2 \sqrt{I_r}}{t} \tag{1}$$

where T^* is time factor; r is the cone radius; I_r is rigidity index, and t is the dissipation time. Teh & Houlsby (1991) have provided theoretical T^* values for various degrees of dissipation (e.g., 20%, 30%, etc.).

Since the calculated c_h should remain consistent regardless of the dissipation degree used, the expressions for c_h at two different degrees of dissipation (assuming constant r and I_r) can be equated. For example, using Eq. (1), for 20% and 50% dissipation:

$$\frac{T_{20}}{t_{20}} = \frac{T_{50}}{t_{50}} \quad \xrightarrow{yields} \quad T_{20} = T_{50}\frac{t_{20}}{t_{50}} \tag{2}$$

Here, T_{20} can be estimated using the t_{20}/t_{50} ratio which can be estimated from the slope of the best-fit line through the origin in a scatterplot of t_{50} versus t_{20} and a T_{50} of 0.245. Other time factors can be estimated similarly using the database approach.

4 RESULTS

4.1 *t_{50} estimation*

The XGBoost regression model trained on basic CPTu parameters (q_t, f_s, u_2, and depth) was unable to reliably predict t_{50} values (Figure 1). The model yielded a high MAPE of 200%, indicating poor prediction accuracy. Furthermore, when comparing predicted t_{50} values to the measured ones in the test set, the points were widely scattered, reinforcing the lack of predictive power. This suggests that the selected input features, although readily available during CPTu testing, do not contain sufficient information to estimate t_{50} with acceptable accuracy using this approach.

In contrast, the LSTM models trained on 20% and 30% pore pressure dissipation levels yielded promising results for predicting t_{50}. As shown in Figures 2 and 3, predicted t_{50} values from both models generally follow the 1:1 reference line, with the 30% model exhibiting reduced scatter and better alignment across a broader range of values.

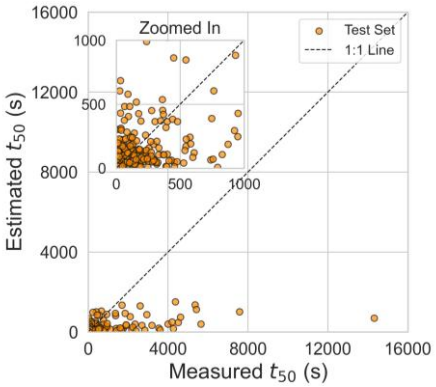

Figure 1. Relationship between measured and estimated t_{50} using XGBoost and basic CPTu parameters.

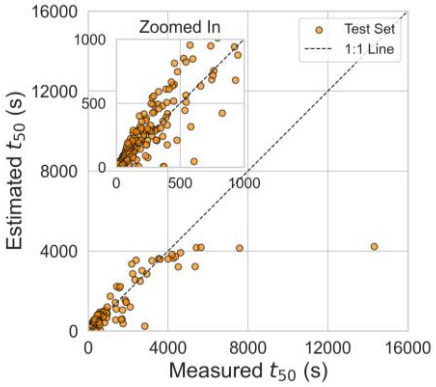

Figure 2. Relationship between measured and estimated t_{50} using LSTM and 20% dissipation data.

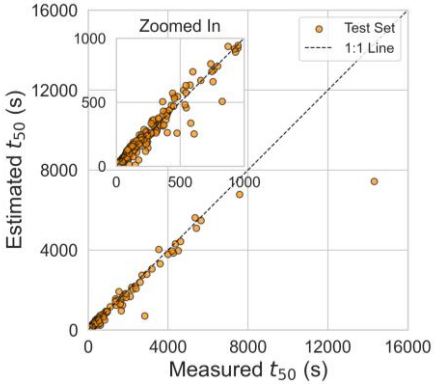

Figure 3. Relationship between measured and estimated t_{50} using LSTM and 30% dissipation data.

The model trained on 20% dissipation data begins to capture meaningful patterns and produced a MAPE of 38%. In comparison, the model trained on 30% dissipation data performed more consistently and achieved a lower MAPE of 20%, indicating a more robust predictive capability. These results suggest that partial dissipation data and ML modelling can potentially be used to estimate t_{50}.

A noticeable decline in regression performance is observed for cases where predicted t_{50} exceeds ~5,000 s when using 20% dissipation data, and beyond ~10,000 s for 30% dissipation data. This drop is most evident in the upper tail of the t_{50} distribution and is primarily due to the limited number of long-duration dissipation tests available for training. As a result, the model struggles to learn and generalize the characteristics of these less frequent, high t_{50} cases. Expanding the dataset to include more long-duration records could potentially enhance the model's performance for high dissipation times. However, the practical impact is minimal since these cases are well beyond drained conditions.

4.2 *Drainage classification*

The confusion matrix in Figure 4 summarizes the classification performance of the ML model trained on basic CPTu parameters to categorize dissipation behavior into Drained, Partially Drained, and Undrained classes. The model achieves an overall accuracy of 63%, with all classes showing almost equal classification performance. Most misclassifications occur between adjacent drainage states (e.g., Drained vs. Partially Drained and Partially Drained vs. Undrained), which aligns with the continuous nature of pore pressure dissipation and overlapping trends in CPTu response. Since class labels were defined using sharp thresholds on measured t_{50} values, some degree of ambiguity near class boundaries is expected and likely contributes to the observed misclassifications. These results indicate that while basic CPTu data carry some signal related to dissipation behavior, more detailed or complementary information may be needed to improve classification accuracy.

As shown in Figure 5, the LSTM model trained on only the first 10 seconds of the pore pressure dissipation time series achieved strong classification performance, with most predictions closely matching the true class labels. The overall accuracy reached 90%. Misclassifications were minimal and primarily occurred between adjacent classes, reflecting the transitional nature of drainage conditions and the difficulty of defining strict class boundaries. These results demonstrate that early-time dissipation curves contain meaningful patterns that the LSTM model can learn to distinguish, even from short segments of the test. This highlights the value of time-series modeling and supports the feasibility of rapid drainage classification based on quick CPTu dissipation response.

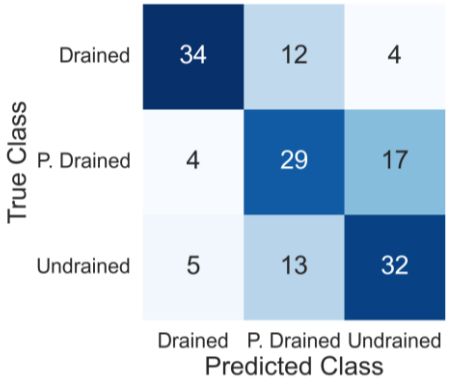

Figure 4. Confusion matrix of XGBoost model trained using basic CPTu data. "P. Drained" stands for Partially Drained.

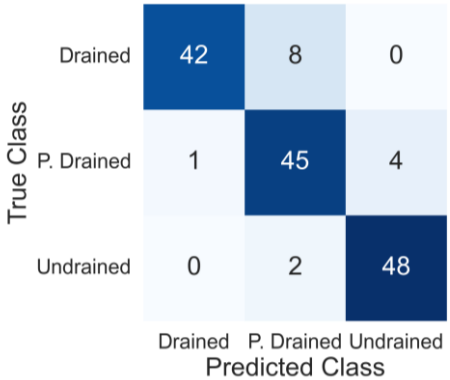

Figure 5. Confusion matrix of LSTM model trained using 10 seconds of dissipation data. "P. Drained" stands for Partially Drained.

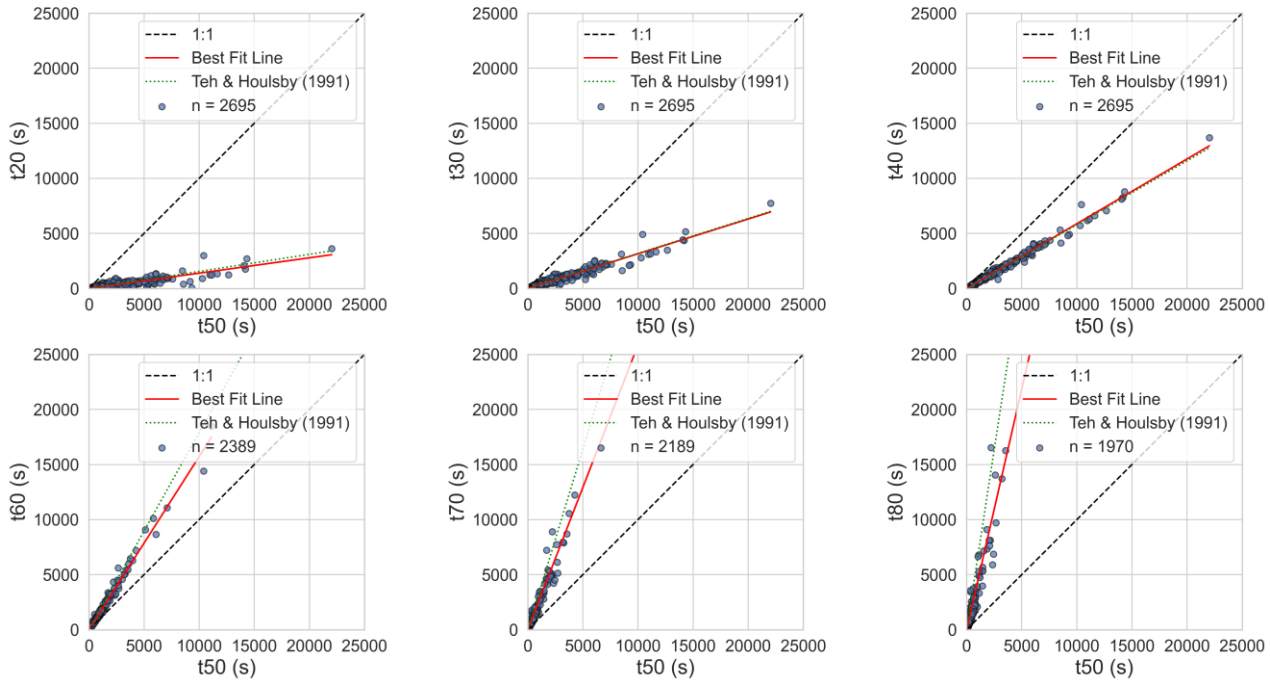

Figure 6. Relationship between t_{50} and t_{20}, t_{30}, t_{40}, t_{60}, t_{70}, and t_{80}.

4.3 *Time factor refinement*

Figure 6 presents the relationships between t_{50} and t_{20}, …, t_{80} using the entire compiled dataset. Each scatter plot includes the best-fit line and the corresponding lines derived from the T^* values proposed by Teh & Houlsby (1991). As shown, there is good agreement between the fitted and theoretical lines at lower dissipation levels (20%, 30%, and 40%), while increasing discrepancies are evident at higher levels (60%, 70%, and 80%). Based on the fitted trends and a reference T_{50} of 0.245, revised T^* values are proposed and summarized in Table 2.

Table 2. Time factor $T^*(u_2)$ versus degree of dissipation.

Dissipation%	This Study	Teh & Houlsby (1991)
20	0.034	0.038
30	0.077	0.078
40	0.144	0.142
50	0.245	0.245
60	0.385	0.439
70	0.634	0.804
80	1.071	1.60

5 CONCLUSIONS

This study presented an ML based approach to estimate t_{50} and to classify soils as drained, partially drained, or undrained. The results showed the challenge of estimating t_{50} purely from basic CPTu parameters. Furthermore, classification of drainage conditions based solely on basic CPTu parameters yielded a limited accuracy of 63%. While the motivation for real-time estimation of t_{50} or drainage condition classification remains strong, the results emphasize the need for additional inputs or alternative modeling strategies to achieve reliable prediction results.

Alternatively, a model trained on 30% dissipation data achieved an average relative error of 20% in estimating t_{50}. This represents a meaningful opportunity to shorten dissipation tests in the field without compromising interpretation quality, thereby improving operational efficiency. Furthermore, an ML model trained on only the first 10 seconds of dissipation data achieved 90% classification accuracy for drainage behavior, suggesting strong potential as a rapid screening tool. In addition, an evaluation of T^* values proposed by Teh & Houlsby (1991) against real world data showed strong agreement at lower dissipation levels but notable deviations at higher degrees of dissipation.

The findings of this study have the potential to enhance the selection of interpretation frameworks, support adaptive testing strategies, and improve the spatial coverage of consolidation estimates. Ultimately, these advances contribute to more efficient and reliable CPTu-based site characterization in complex soil profiles.

6 REFERENCES

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August, 785–794.

DeJong, J. T., and Randolph, M. 2012. Influence of partial consolidation during cone penetration on estimated soil behavior type and pore pressure dissipation measurements. *J. Geotech. Geoenviron. Eng.*, 138(7), 828–841.

DeJong, J. T., Jaeger, R. A., Boulanger, R. W., Randolph, M. F., and Wahl, D. A. J. 2012. Variable penetration rate cone testing for characterization of intermediate soils. *Geotechnical Site Characterization 4*, 25–42. London: Taylor & Francis.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural Computation, 9(8): 1735–1780.

Robertson, P. K. 2016. Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — an update. *Can. Geotech. J.*, 53(12), 1910–1927.

Teh, C. I., and Houlsby, G. T. 1991. An analytical study of cone penetration test in clay. *Géotechnique*, 41(1), 17–34.